International Journal of Assessment Tools in Education
2023, Vol. 10, No. 2, 257–278
https://doi.org/10.21449/ijate.1210479
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Application of the Rasch model in streamlining an instrument measuring depression among college students

Sherwin Balbuena [1,*]

[1]Dr. Emilio B. Espinosa Sr. Memorial State College of Agriculture and Technology, Masbate, Philippines

ARTICLE HISTORY

Received: Nov. 26, 2022
Revised: May 14, 2023
Accepted: May 14, 2023

Keywords:
Rasch,
Depression,
Precision,
Measurement,
USDI.

Abstract: Depression is a latent characteristic that is measured through self-reported or clinician-mediated instruments such as scales and inventories. The precision of depression estimates largely depends on the validity of the items used and on the truthfulness of people responding to these items. The existing methodology in instrumentation based on a factor-analytic approach has limited applicability, especially in the detection of sources of measurement error in item- and person-level analyses. While there are probabilistic approaches such as the use of Item Response Theory and the Rasch model in validating instruments, there are no definite guidelines on the sequence of steps to follow. This study explored the suitability of the Rasch model in assessing and streamlining the University Student Depression Inventory (USDI) using a sequential strategy based on the item response model assumptions, which involves fitting the data to the model through the elimination of misfits, analyzing retained items, and constructing measures. The strategy was applied to two sets of survey data collected from the same population of college students enrolled in a Philippine university but in different semesters. Results showed that the Rasch procedure was able to detect misfit items and persons, which guided decisions regarding the removal of problematic items and persons while preserving the reliability of the original scale. The methodology used was found to be replicable, as the analyses for the two datasets yielded comparable results in terms of number of items retained, item estimates and severity ordering, and distribution of student depression measures.

1. INTRODUCTION

Latent characteristics are human traits, constructs, or attributes that are neither directly observable nor tangible (e.g., feelings, affect, intelligence, etc.). As such, it is made manifest by eliciting responses from human subjects through interviews, tests, or self-reports. Usually, an individual responds to questions or items, and his/her responses are considered outer or observable indicators of this inner but unobservable human condition. So, the existence and quantity of the latent characteristic must be deduced from those observed, manifest responses (Bond & Fox, 2013).

In practice, a latent characteristic is indirectly measured through the administration of instruments. Prior to administration, these instruments undergo a rigorous development and

*CONTACT: Sherwin Balbuena balbuenasherwine@debesmscat.edu.ph Dr. Emilio B. Espinosa Sr. Memorial State College of Agriculture and Technology, Masbate, Philippines

e-ISSN: 2148-7456 /© **IJATE** 2023

validation process. At the early stages of the process, an item pool was developed based on the concept deduced from an underlying construct or latent characteristic, and each item was designed to capture information about the construct. Then the initial pool of items was piloted with a sample of the target population. The responses of this sample are usually subjected to factor analysis to assess the psychometric properties of the items. The items that load highly on the factors extracted based on eigenvalues (amounts of variance contributed) are retained and are therefore chosen to make up a scale. This process is based on a factor-analytic (FA) approach.

The FA method, also known as Classical Test Theory (CTT), is based on the premise that a test is valid and reliable if it comprises items that have high loadings with known variables related to the latent trait being assessed and if the responses to given items are consistent. Due to the presence of unpredictable items and person responses, it is more error-prone since it does not offer item- and person-level metrics to identify items that did not function as planned and to detect response sets. Furthermore, the estimation of individual abilities is test-dependent, while the estimation of item difficulties and discrimination is sample-dependent (Kohli et al., 2015).

Probabilistic approaches to instrument item analysis have emerged in recent years. One example is the use of item response theory (IRT) or the Rasch model (RM) to investigate the psychometric properties of constructed items and to validate and refine existing tests, questionnaires, or scales. Using this new approach, the response of a person to an item is modeled using a logistic function relating the person's underlying ability and the item's difficulty. As applied in scaling, an initial pool of items is created based on the underlying latent characteristic. The items are designed to gather information about the attribute at different severities in the latent continuum (a linear scale where a person can be identified as having less or more of the attribute). Next, the constructed scale is administered to a sample from the intended population. The responses of the sample are analyzed for model fit using fit statistics for both items and persons, and those items that do not deviate significantly from model expectations are retained in the new scale. This is one advantage of IRT, since it provides sample-free and test-free measures by estimating item and person parameters separately using conditional maximum likelihood methods and by requiring that response data fit the model.

The amount of the latent characteristic is quantified based on the outcome of the person's response to instruments. Measures of the latent characteristic of interest are obtained by summing up the person's ordinal item responses. The adequacy of the quantification of latent characteristics based on raw scores depend on the length of the questionnaire. In the factor-analytic point of view, the more items you include, the more valid and reliable the instrument becomes, as it gathers more information about the latent concept of interest. However, instruments with high reliability indices may contain redundant items (Boyle, 1985). Furthermore, using longer questionnaires would increase respondent burden, which would subsequently lead to low response rates (Stanton et al., 2002) and poor data quality (Galesic and Bosnjak, 2009; Maloney et al., 2011). More importantly, the validity of the data largely depends on the truthfulness of persons responding to questionnaire items and on the appropriateness of the items included in the questionnaire. Analysis of flawed data due to invalid responses and items would certainly produce invalid results and inferences about the latent characteristic.

An example of a latent characteristic that is currently attracting attention is depression. Depression is a common mental health problem that can impair an individual's functioning at home, work, or school (WHO, 2017). It is a medical condition characterized by a set of behavioral, cognitive, social, and biological symptoms (Hyde et al., 2008). Its severity ranges from a mild feeling of sadness to serious suicidal thoughts (Olsen et al., 2003; Forkmann et al., 2013; Balsamo et al., 2014). Depressive symptoms often manifest even at an early age, can be

recurrent, and, if left untreated, will lead to the development of severe mental disorders (Hankin, 2006). Diagnosing the early signs of depression and providing appropriate interventions (e.g., counseling) can potentially prevent the progression of the disease, which is less costly than treating patients with severe depression (O'Connell et al., 2009).

The diagnosis of depression relies on self-reported instruments and diagnostic interviews, where the presence of a sufficient number of symptoms qualifies an individual as depressed. Over the past five decades, various depression inventories such as Beck's Depression Inventory (BDI), Patient Health Questionnaire (PHQ), and others have been developed, validated, and used to assess depression levels. These inventories have been employed in both general and specific patient populations, providing valid measures of depression. Typically, the patient's responses are analyzed using predetermined cutoff points to determine depression presence and severity. Different measurement frameworks, including CTT and IRT, have been employed to estimate depression levels using these instruments, yielding comparable measures (e.g., Stansbury et al., 2006; Shea et al., 2009; Balsamo et al., 2014; Wongpakaran et al., 2019).

The prevalence of depressive disorder is estimated using either self-reported instruments or a clinician-rating scales. The use of self-reported instruments has been found to overestimate the proportion of depressed individuals measured by diagnostic interviews. The lower point prevalence of depression obtained in diagnostic interviews might be due to the stringent criteria being adopted by clinicians in screening depressed individuals and could be associated with the socio-demographic characteristics of patients. Hence, the use of both methods in estimating depression prevalence was recommended (Lim et al., 2018).

Obtaining measures of depression is important for informing clinicians or researchers about this latent disease and its prevalence. However, measuring depression can be difficult, as it is done through a series of thorough observations and interviews with the patient. In the absence of psychiatric experts or clinicians to confirm the presence or absence of the disease, treatment is sometimes delayed, and undetected mild cases progress to severe cases. Alternative sources of depression measures are needed to detect not only the severe cases but also those at risk and provide a more inclusive mental health assessment. There are many available depression scales, but most of them are lengthy. We need to provide a rigorous statistical methodology by which we can assess and streamline these scales to efficiently measure depression for clinical use and research purposes.

To date, procedures for instrument assessment using IRT have been varied across fields of inquiry. In many health studies, using the Rasch model in instrument short-form development and psychometric validation is referred to as Rasch analysis. The analysis involves testing the following: (a) the data’s fit to the model; (b) the appropriateness of response format for polytomous items; (c) differential item functioning; (d) targeting of persons and items; (e) reliability; (f) local independence; and (g) unidimensionality (Tennant & Conaghan, 2007). Unfortunately, many studies applying Rasch analysis did not provide a definite sequence of steps to follow in assessing item properties, which could be used by researchers as a guide in streamlining existing instruments. Hence, there is a need to develop a robust statistical procedure for instrument quality assessment using Rasch analysis.

The general objective of this research was to evaluate the applicability of the Rasch model for assessing and analyzing an instrument measuring student depression for possible item reduction without compromising validity and reliability.

The specific objectives were as follows:

1. To determine the suitability of the Rasch model in the construction of scales for measuring depression in students;

2. To streamline the questionnaire items given the same precision level as that of the original instrument;
3. To develop an appropriate procedure for analyzing questionnaire items, which is applicable in evaluating the quality of instruments measuring other latent characteristics; and
4. To test the replicability of the procedure when applied to two datasets derived from the same population.

1.1. The Rasch Model

This study used the following assumptions of the Rasch model in forming the basis for deciding which items to discard, retain or review: unidimensionality (assessed using Rasch fit statistics and Martin-Löf test), local stochastic independence (detected using item residual correlations), and no differential item functioning (determined using ordinal logistic regression in IRT with gender as the only reference group). Person fit was also investigated to identify persons with aberrant response patterns to be excluded from the analysis. The procedure developed in this study was empirically applied to a dataset derived from two surveys on the mental health of Filipino college students. However, it is assumed that this procedure will be sufficiently robust when applied to streamlining instruments measuring other types of latent characteristics.

Named after its originator Georg Rasch (1960), a Danish mathematician, the Rasch model is a statistical model that is used to analyze data from educational and psychological measurement instruments, such as tests, surveys, and questionnaires. It is a type of item response theory model, which is a framework for understanding how individuals respond to specific items on a measurement instrument. The Rasch model is based on the assumption that the difficulty of an item and the ability of the person responding to the item are related in a specific way. The model specifies a mathematical relationship between the two, which allows researchers to estimate an individual's ability level based on their responses to a series of items (Bond & Fox, 2013). The idea of invariant measurement, which is developed through specific objectivity, governs the model. Item difficulty may be evaluated independently of the people included in the sample, and individual ability can be estimated irrespective of the test items (Wright & Linacre, 1987).

One feature of the Rasch model is that it provides estimates of person measure (or person ability) and item location (or item difficulty). The item and person estimates can be calibrated on the same logit scale. Furthermore, unlike traditional ordinal scales using unequal-spaced intervals of scores to rank the underlying ability from less to more, Rasch scaling permits equal spacing of intervals (Yu, 2011). This is done by finding the logarithm of the original (raw score-based) scaling used. Theoretically, the difference between two discrete raw scores (which are actually sums of affirmative or ordinal item responses) is not meaningful. Hence, Rasch scaling resolves this problem by converting discrete raw scores into continuous logit measures after fitting the data to the model, which can be used not only to determine who has more or less of the ability but also to compare the relative distances between person abilities.

Essentially, the Rasch model is based on the theory about the construct of latent characteristics under scrutiny. The construction of items is guided by a thorough understanding and definition of the latent concepts involved and of the behavioral manifestations (i.e., responses) that represent the construct. The order of item locations can be empirically ascertained by estimating the item difficulty parameters after fitting the response data to the Rasch model.

1.2. Measuring Depression in Students

Detection of depressive disorder in an individual is done using self-reported instruments and diagnostic interviews, where the presence of a sufficient number of symptoms qualifies the individual as depressed. Diagnosis of depression is also done through the analysis of patient's self-reported symptomatology in interviews. For the past five decades, several clinical or research depression inventories have been developed, validated, and then used to come up with

measures of the depression level. Some inventories commonly used for the general population include Beck's Depression Inventory (BDI; Beck et al., 1961), Patient Health Questionnaire (PHQ; Spitzer et al., 1999), Hospital Anxiety and Depression Scale (HADS; Zigmond and Snaith, 1983), Depression Anxiety Stress Scale (DASS; Lovibond and Lovibond, 1995), Centre for Epidemiological Studies - Depression Scale (CES-D; Radloff, 1977), Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960), and Zung Self-Rating Depression Scale (Zung, 1965). Some of these inventories have already been used in specific patient populations. Disease-specific depression scales, which are new versions of the above inventories, have been administered and found to provide valid measures of disease-related depression levels.

An instrument called the University Student Depression Inventory (USDI; Khawaja and Bryden, 2006) is used to screen for student depression, a non-disease-related kind of depression. The measure was initially created and tested on a sample of university students from Australia. Following principal component factor analysis using oblique and orthogonal rotation techniques, the developers of USDI identified three (3) factors from an original set of 125 produced items. The final instrument (found in Appendix A) was composed of 30 items classified into three sub-scales: lethargy (LG; 9 items), cognitive-emotional (CE; 14 items), and academic motivation (AM; 7 items). To obtain a measure of student depression using the USDI, the scores in the instrument are added, and a higher total is interpreted as an indication of a higher level of depressive disorder. Recent local studies used the USDI to determine some factors associated with depressive symptoms among Filipino college students (Lee et al., 2013; Lailo, 2018). However, this instrument alone cannot diagnose or confirm that a student actually has depression and should be used only to measure the degree of vulnerability of a student to developing severe depression (Lailo, 2018).

The validity and reliability of the USDI have already been established. Further psychometric validation studies were conducted to confirm the factor structure of the USDI using multi-cultural student populations (Sharif et al., 2011; Romaniuk & Khawaja, 2013; Khawaja et al., 2013; Habibi et al., 2014). The instrument has already been used in a number of studies estimating the prevalence of depressive symptoms among tertiary students (e.g., Mikolajczyk et al., 2008; Deb et al., 2016; Gesinde & Sanu, 2014). However, research on the application of Rasch analysis in validating the USDI was limited.

2. METHOD

2.1. Research Design

To achieve the objectives of this study, an exploratory type of research design was used. The theory on which the Rasch model is based is that the items in a scale or questionnaire were constructed with varying levels of difficulty, which can be ordered along a single continuum. The suitability of this model in analyzing questionnaire items (with unknown difficulty levels) was explored through the application of the model to the response data obtained from university-wide mental health surveys. In light of the Rasch model assumptions, a strategy for streamlining questionnaires was developed and applied to the data set to explore its soundness and replicability.

2.2. Data Sources

The data used in this study are the responses of college students to items in the University Student Depression Inventory (USDI). This questionnaire was used on two mental health surveys in a state university in Southern Luzon to detect the presence of depression in students as manifested in depressive symptoms. The first dataset was taken from the result of the university-wide survey conducted in the school year 2018-2019, referred to here as the STAT 173 survey (UPLB INSTAT, 2018), which involved 441 college students. The survey aimed to determine the level of depression among undergraduate students, specifically to describe

depression incidence among students and to identify possible determinants of student depression. The second dataset was taken from the result of the survey conducted by Lailo (2018) in the school year 2017-2018 at the same university, which involved 169 college students.

2.3. Methodology

By adhering to the assumptions of the Rasch model, the following steps were done to achieve the objectives of the study: (1) select an appropriate model; (2) estimate the model parameters and identify misfitting persons; (3) after removal of person misfits, re-estimate the model parameters and identify misfitting items; (4) after removal of item misfits, re-estimate the model parameters and identify misfitting items until no further items are misfitting; (5) assess the reliability at each instance of item/person removal until reliability declines tremendously; (6) order the item severity estimates and check for consistency with established symptomatology to assess construct validity; (7) detect local dependence (LD) and differential item functioning (DIF) for possible item redundancy and bias; (8) estimate person measures and transmute with raw scores; and (9) locate the thresholds for varying severity levels. The analyses involved in the procedure were implemented in R, Microsoft Excel, and SPSS Version 28.

The same strategy described above was applied to the analysis of Lailo's (2018) data. The response data were obtained using the same instrument and from the same population of college undergraduate students, but in different semesters. The resulting streamlined versions of the instrument from the analyses of the two data were compared in terms of the number and similarity of retained items, overall instrument reliability, and constructed measures of student depression.

The main purpose of this study was to develop a strategy based on Rasch model tests for assumptions to assess and streamline a depression scale and produce a shorter version of the original scale, which is equally valid and reliable. Using two sets of survey data, the soundness of the strategy was determined as will be described in this section.

3. FINDINGS

3.1. Analysis of STAT 173 Data

3.1.1. *Model selection*

The Rasch model is a family of statistical models used in psychometrics for analyzing categorical data. The selection of the appropriate Rasch model depends on the type of data being analyzed. In general, there are two main types of Rasch models: dichotomous and polytomous. The dichotomous Rasch model is designed for binary data, where responses are either correct or incorrect. The polytomous Rasch model, on the other hand, is designed for data with more than two response options, such as Likert scales. Therefore, choosing the most appropriate model depends on the nature of the data and the research question at hand.

The most basic formulation is the dichotomous Rasch model (DRM), which is also referred to as the one-parameter logistic model (1PL). Let $X_{ni} = x \in \{0,1\}$ be a dichotomous random variable, where x = 0 and x = 1 indicate "no" and "yes" responses, respectively, to a questionnaire item. The function

$$P(X_{ni} = 1) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{1}$$

models the probability that person n will agree with item i, where δ_i is the difficulty of item i and β_n is the ability of person n. This function conjectures that the higher a person's ability

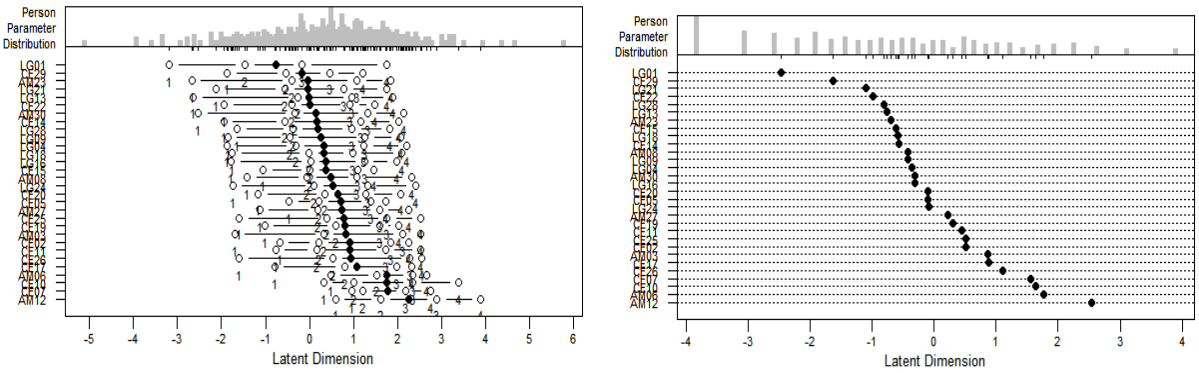
relative to the difficulty of an item, the higher the probability of an affirmative response on that item, a relation that can be illustrated by a sigmoid graph with the person's ability as the abscissa and the probability of agreeing as the ordinate.

On the other hand, the Partial Credit Model (PCM; Master's 1982) is a type of Rasch model that is used to analyze polytomous data, which is data that has more than two response categories. This model assumes that a person's response to an item is a function of the person's ability, the item's difficulty, and the threshold parameters that describe the level of difficulty at which the person is able to transition from one response category to the next.

Using the two models DRM and PCM, two sets of parameter estimates were derived using the eRm package (Mair & Hatzinger, 2007) in R. Results showed that the least severe and the most severe items for both models were LG01 ("I am more tired than I used to be") and CE20 ("Going to university is pointless"), respectively. Although the item severity orders were found to be similar (with minor changes in ordering for items with moderate severity measures that are close to one another), the infit and outfit indices changed remarkably after the response was dichotomized.

The person-item map (PIM) also provides rich information about the relationships between item and person estimates and their distributions. Figure 1a shows the nearly symmetric person distribution after fitting PCM, with central tendencies located between 0.0 and +2.0 logits. The item severity range (-0.77 to +2.25) along the latent dimension also spanned the width of the person distribution, which means that the USDI was well-targeted for the given population of college students. When data were fitted to DRM, the person distribution became more dispersed and the item severity range increased in width (-2.47 to +2.54 logits) as shown in Figure 1b, although the item severity ordering did not change significantly.

Figure 1. *STAT 173 data PIMs under (a) PCM and (b) DRM.*



(a) (b)

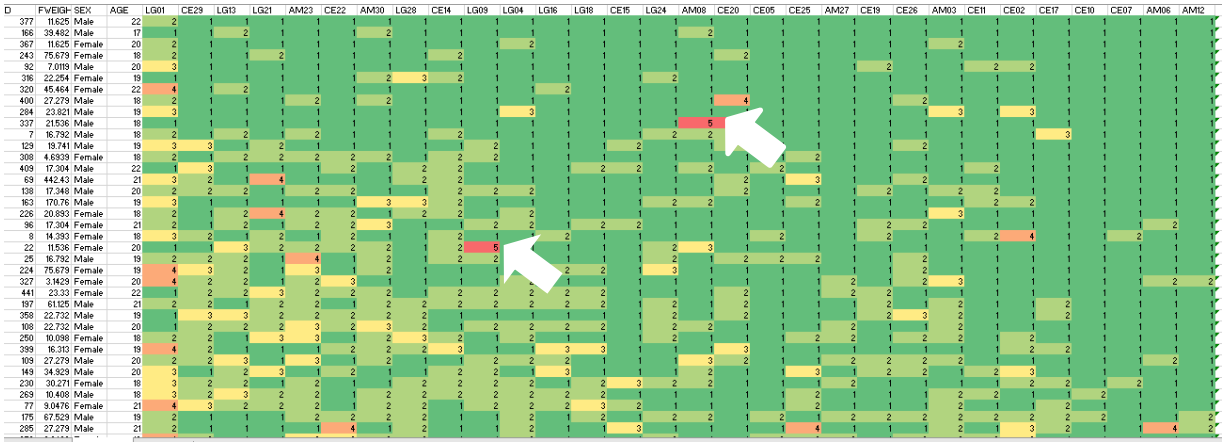
The alteration of the person measure distribution above suggests the inappropriateness of dichotomizing USDI response data. Hence, it is logical to discard the use of DRM as an option to modelling originally polytomous data. Hence, the polytomous PCM was used to analyze the STAT 173 data. PCM was also used in the analysis of STAT 173 data to warrant comparison of results with Lailo's data since it was found out that a 5-point Likert scale did not apply for some items in the latter.

3.1.2. *Person misfits*

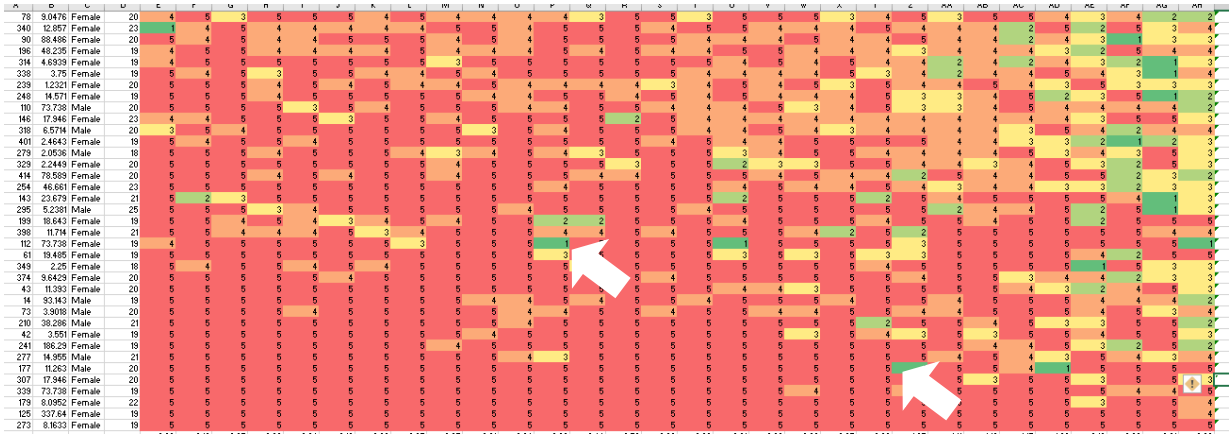
After fitting the PCM, person estimates were obtained with corresponding person fit analysis. The eRm package provides a summary of fit analysis with chi-square, outfit/infit mean square, and outfit/infit t indices. From the result of fit analysis, there were 69 persons with very high values for both infit and outfit exceeding the threshold of 1.3 and with chi-square p-values less than 0.05; hence, they were labelled as misfits (also referred to as underfit persons). These

persons have so highly unpredictable responses that they distort the measurement system (Linacre, 2002), or they are known to cause measurement error. Another helpful approach to detecting misfits is by arranging the items and persons according to their estimated locations in the order from less to more severe/depressed and then by examining their patterns of responses to items. To illustrate this approach, a Guttman (1950) scalogram heatmap showing the noticeable patterns of responses by misfit persons was constructed as shown in Figure 2. Unusual patterns of responses, as represented by uncommon cell colors, also represent a substantial deviation from the expectations of the Rasch model (i.e., large value of residual) as was detected by infit and outfit statistics. For example, the two responses in Figure 2a (shown by the arrows) represent the high-category responses of less depressed Person #337 to items AM08 and CE20 and the high-category response of less depressed Person #22 to Item LG09. Corresponding outfit(infit) for these persons show very high mean square values, 5.62(5.05) and 1.67(2.12), respectively. For more depressed persons #112 and #177 in Figure 2b, these unusual response patterns were also detected by high outfit(infit) statistics, 3.23(3.11) and 2.68(3.63), respectively.

Figure 2. *An Excel®-generated Guttman scalogram heatmap showing the patterns of item responses of (a) low-ability (less depressed) and (b) high-ability (more depressed) persons. A row containing totally different or contrasting cell colors (e.g., a red surrounded by a dominant green or a green surrounded by a dominant red) represents the response of a misfitting person.*



(a)



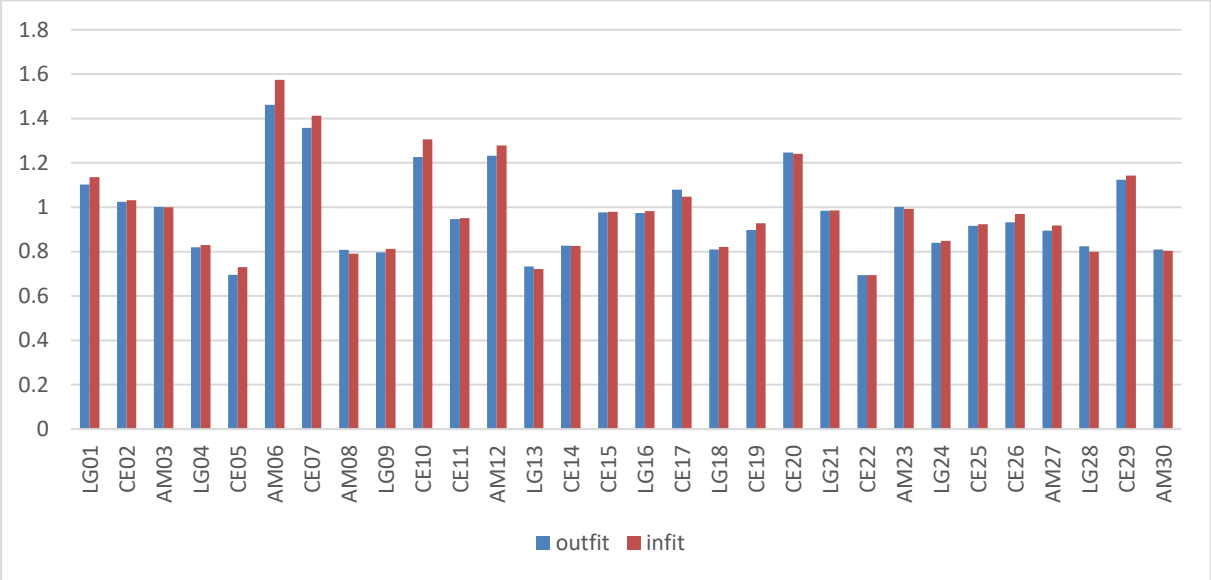
(b)

These misfitting persons were removed from the analysis, as the information they provided did not contribute to useful measurement of student depression. Following this weeding out, a 16% decrease in the number of samples was noted.

3.1.3. *Item misfits*

Using the new dataset with 372 fit persons, the item and person parameters were estimated using the PCM. The new results were obtained, which include the estimates when the complete sample was used. Comparing the two results, slight changes in the item severity estimates and ordering and noticeable changes in outfit/infit for some items were observed. Item AM06 ("I don't attend lectures as much as I used to") and CE07 ("I have thought about killing myself") consistently had very high values for both infit and outfit; hence, they were labelled as misfits. Item CE10 ("No one cares about me") had an infit value exceeding the threshold, but this is considered trivial. Hence, only two items (AM06 and CE07) were considered for removal, as these items were believed to contribute substantial error variance to analysis.

Figure 3. *Clustered column bar charts for the infit and outfit of USDI items in STAT 173 data after removal of 69 misfitting persons.*



Another approach in assessing fit of the items is by examining infit-outfit relationship through graphical method using clustered column bar (CCB) charts. In Figure 3, the unusual patterns of infit and outfit values can be observed, and problematic items can be identified. At a glance, one can note that items AM06 and CE07 have remarkably high outfit-infit values compared to the rest of the items. This observation corroborates the previous decision made to discard the two items, which are believed to be unproductive for the construction of measures for student depression.

After removal of items AM06 and CE07, three more items were found to show very high outfit-infit values: AM12 ("Going to university is pointless"), CE10 ("No one cares about me"), and CE20 ("I spend more time alone than I used to"). This is to be expected, since these items previously had tall outfit-infit bars in Figure 3 secondary to those of the two already discarded, meaning they contribute substantial measurement errors. Following item removal, both outfit and infit values of these items escalated and exceeded the cutoff, hence they were labelled as misfit items. Further removing these three items from the data showed no further items misfitting the PCM. In Table 1, there are no items with remarkably high fit statistics. This means that the STAT 173 data with the remaining 25 items conformed to the unidimensionality assumption (Wright and Panchapakesan, 1969). Furthermore, Martin-Löf test showed a nonsignificant result (LR = 857.936, df = 24, $p > 0.05$), indicating that the data appears to be unidimensional.

Table 1. *STAT 173 data item estimates and fit statistics after removal of five misfitting items, showing satisfactory fit statistics.*

Item	Severity	Outfit	Infit
LG01	-0.61	1.12	1.15
CE02	1.56	1.21	1.21
AM03	1.53	1.12	1.14
LG04	0.69	0.87	0.88
CE05	1.30	0.79	0.81
AM08	1.04	0.87	0.85
LG09	0.65	0.83	0.85
CE11	1.56	1.11	1.09
LG13	0.27	0.75	0.74
CE14	0.60	0.93	0.90
CE15	0.86	1.06	1.03
LG16	0.74	1.06	1.07
CE17	1.70	1.21	1.17
LG18	0.78	0.86	0.86
CE19	1.37	0.99	1.03
LG21	0.33	1.05	1.04
CE22	0.47	0.70	0.69
AM23	0.37	1.03	1.03
LG24	0.99	0.89	0.90
CE25	1.35	1.01	1.02
CE26	1.48	1.00	1.05
AM27	1.35	0.95	0.98
LG28	0.58	0.82	0.80
CE29	0.09	1.16	1.16
AM30	0.57	0.82	0.81

3.1.4. *Reliability assessment*

Since the first PCM estimation and throughout the instrument assessment process, reliability analysis has been carried out, especially when a group of misfit respondents or a group of misfit items were excluded from the analysis. The reliability analysis summary for each instance of excluded persons or items is shown in Table 2. The reliability indices in Stage 1 were found to be very close to Cronbach's alpha. Note that even after the elimination of 69 people, the PSI remained constant. The reliability was unaffected significantly by subsequent item reductions either. Additionally, the person separation is 5.69, which indicates that the sample of college students may be divided into around six distinct depression severity categories. This suggests that the new 25-item USDI is as equally precise and accurate in measuring student depression as the longer USDI.

Table 2. *Summary of person separation reliability assessment at each stage of fit analysis and item/person reduction of STAT 173 data.*

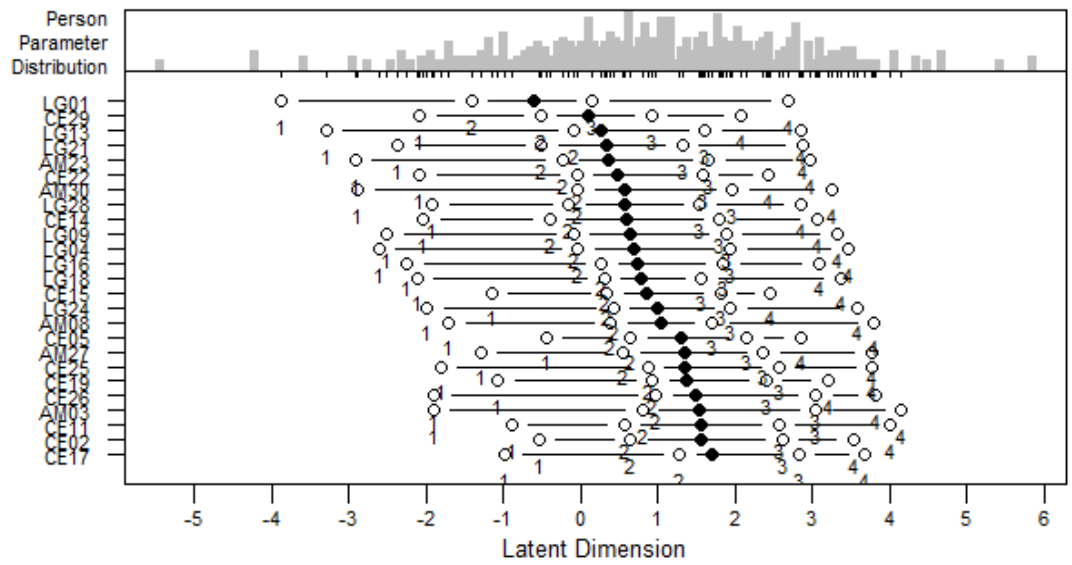
Stage	Instance	Observed variance $\sigma^2_{\hat{\beta}}$	Model error variance $\sigma^2_{\hat{e}}$	Reliability (PSI)
1	Used 30 items and 441 persons*	2.0097	0.0636	0.97
2	Used 30 items and 372 fit persons	2.6687	0.0737	0.97
3	Used 25 items (AM06, CE07, AM12, CE10, and CE20 discarded) and 372 persons**	3.1641	0.0894	0.97

*Cronbach's alpha = 0.969; **Person separation = 5.69

3.1.5. *Item severity ordering*

The item severity estimates are regarded as reliable indicators of location when a person falls on the severity spectrum of depression provided that all the items fit the PCM. The relative ordering of the 25 remaining items based on the values of item estimations represented in logits may be determined by creating a PIM. Based on the item location (solid circle) and threshold (hollow circle) estimates, the PIM in Figure 4 shows how the items are arranged from less severe to more severe manifestations of student depression. It is clear that the majority of the items in the below-median group fall under the LG subscale, whereas the majority of the items in the above-median group belong to the CE subscale, if the items are divided into two groups (i.e., less severe and more severe) based on their location above or below the median item measure. The AM subscale items do not display any specific severity classification. Based on extreme locations, anhedonia, or the loss of interest in previously enjoyed activities (cognitive-emotional), as manifested in Item CE17 ("The activities I used to enjoy"), is the most severe symptom of depression. Anhedonia is actually one of the primary signs of major depression (American Psychiatric Association, 2013).

Figure 4. *The PIM for the remaining 25 USDI items in STAT 173 data showing the locations of the items along the latent dimension (depression severity) with corresponding response category threshold estimates.*



3.1.6. *Detection of LD*

Using the criteria of Smith (2000) and Chen and Thissen (1997) on item residual correlations, one pair of LD items was identified. With a correlation coefficient of 0.3151, Items CE25 ("I

feel withdrawn when I'm around others") and item CE26 ("I do not cope well") were found to be locally dependent. The choice in this case is to examine each item's contents for potential revision rather than to discard one of the items. According to research, LD items should be revised by combining the two items into a single "super-item" (Wainer & Kiely, 1987).

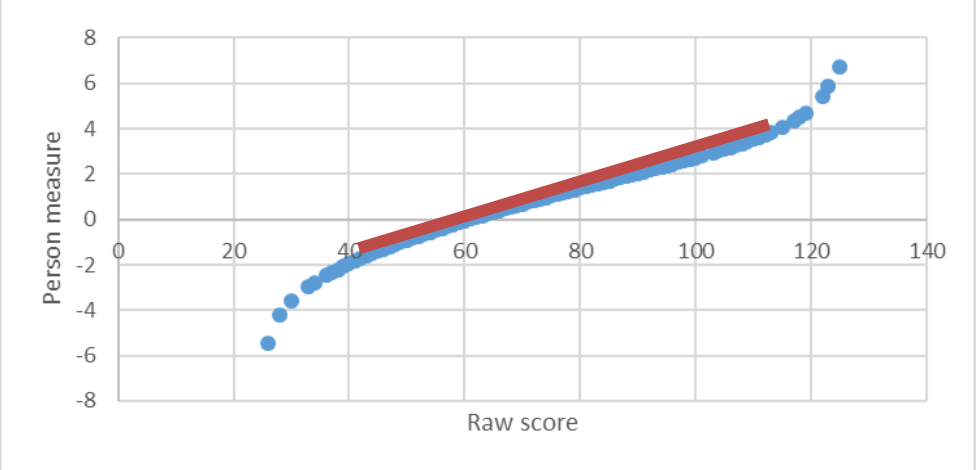
3.1.7. *Detection of item bias or DIF*

Using ordinal logistic regression in IRT with gender as the only reference group and then following the chi-square criterion based on the likelihood ratio χ^2 test (Swaminathan and Rogers, 1990), three items (LG01, AM03, and AM08) were marked for DIF. It is recommended that these items be retained and check their content for any idiosyncratic meanings or getting gender-specific item attributes for these items, which may be utilized to establish different norms for male and female students.

3.1.8. *Construction of measures*

By adding the ordinal values assigned to each student's categorical responses across all items for the remaining 25 items in the streamlined USDI, the level of depression is calculated for each student. The result is a number called the raw score, which has a range of 25 to 125. Equivalent interval-level measures for the raw scores were obtained because all the items fit the PCM. For some middle scores, an almost linear relationship can be seen (see Figure 5) if the scatterplot of raw scores and Rasch person measures is constructed. The scatter plot resembles a straight line between raw scores 40 and 110. Within this range, valid transformations from discrete raw score to continuous person measure are offered by interpolation using a linear function.

Figure 5. *Raw score to Rasch measure transformation scatterplot in STAT 173 data, showing the almost linear relationship between raw scores 40 and 110.*



3.1.9. *Scaling and classifications of depression level*

To determine provisional thresholds for classifying students into groups with varying levels of depression (i.e., very high, high, moderate, and low), the averages of item thresholds of all the items were considered. This approach was considered valid since the USDI instrument was well-targeted for the given population of college students, because it was discovered that the distribution of item measures and person measures were similar. Additionally, no item was found to have disordered thresholds, indicating that the scale structure used was effective and that student responses increased monotonically as depression levels increased. Hence, the thresholds estimated from these scale structures and probable responses to items representing depressive symptoms could be used to demarcate various levels of depression experienced by students.

Table 3. *Current classification thresholds for each category in STAT 173 based on continuous person measures expressed in logits with corresponding discrete raw score thresholds.*

Category	Person Measure		Raw Score	
	Lower limit	Upper limit	Lower limit	Upper limit
Low	—∞	-1.95	25	40
Moderate	-1.95	0.64	41	69
High	0.64	3.24	70	106
Very high	3.24	+∞	107	125

While the person separation in Table 2 suggested six categories of depression severity, this study used only four categories to warrant comparison with the results of the original STAT 173 data analysis using four levels of depression severity (i.e., 30-59 low, 60-89 moderate, 90-119 high, 120-150 very high), referred to here as the previous scale. To determine the cutoff for each category, the average of threshold estimates for categories 1 and 4 were computed and found to be at -1.95 and +3.24 logits, respectively. These two cutoff points represent the upper limit for the low depression category and the lower limit for the very high depression category. The middle threshold (or cutoff for classifying between moderate and high levels) was determined by finding the midpoint between the average threshold estimates for categories 1 and 4, which is +0.645 logits. Hence, the common distance between thresholds 1 and 2 and between thresholds 2 and 3 is 2.595 logits. The classification cutoffs (referred to as the current scale) are shown in Table 3 in which the equivalent raw score category limits are also indicated.

Using the data without misfit persons and items, the number of persons with inconsistent classifications between the previous and current scales was determined. Results showed that 90 out of 372 persons had inconsistent classifications, majority of which were transferred from a lower category to the next higher category. The inconsistencies occurred as a result of the change in raw score intervals when the logit-based thresholds were used; that is, some category interval widths were widened or narrowed when the equivalent raw score category limits were used as cutoffs. But the current classification based on interval-level logit measures is more valid, since the previous one was based on intervals of discrete raw scores derived from the sum of ordinal response data for which equally spaced intervals cannot be constructed (Yu, 2011).

3.2. Analysis of Lailo's Data

After applying the same sequential strategy on the analysis of Lailo's data, the following results were obtained. Using the outfit and infit statistics based on PCM, 17 out of 135 persons were found to be misfits and then removed. Five items (AM06, CE07, LG18, CE20, and AM23) were identified as misfits and then removed. Following this removal of sources of measurement error, the remaining 25 items achieved a good fit to the PCM while preserving the internal consistency of the original USDI at 0.96 PSI. Furthermore, no items were detected for gender DIF, while some adjacent questionnaire items were flagged for LD. Locally dependent items were not considered for removal as the result of item residual correlations might have been caused by respondents' acquiescence to the redundancy of questionnaire items or simply a false positive detection.

The distributions of the estimated person and item locations showed that the USDI was well-targeted to the given population of college students. Measures of depression at various severity levels were also constructed based on the estimated thresholds as shown in Table 4.

Table 4. *Current classification thresholds for each category in Lailo's data based on continuous person measures expressed in logits with corresponding discrete raw score thresholds.*

Category	Person Measure		Raw Score	
	Lower limit	Upper limit	Lower limit	Upper limit
Low	−∞	-1.13	25	43
Moderate	-1.13	+0.39	44	68
High	+0.39	+1.91	69	100
Very high	+1.91	+∞	101	125

Using the previous classification thresholds for Lailo's data (the same ordinal scale used in STAT 173 data) and the new classification thresholds, 11 persons were found to be inconsistently classified, all of whom were transferred to a next higher category.

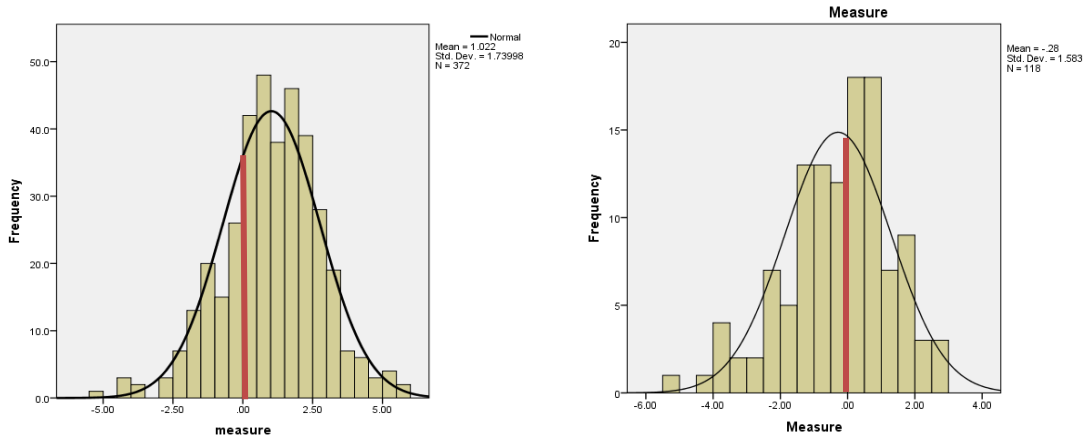
3.3. Results for STAT 173 and Lailo's Data: A Comparison

From the above analyses, evidences on the properties of the USDI items were gathered, which served as basis for item removal and further item analysis. Removal of items was based on misfit values only. Items flagged for LD and DIF were retained for further review of item contents. There were five items removed, but only three of these items were common to both data contexts (AM06, CE07, and CE20). Separate orderings of the item location estimates of the retained items common to both data contexts showed high Spearman rank-correlation coefficient ($\rho = 0.87$, $p < 0.05$), which indicates identical orderings of item severities. Finally, the reliability of the instrument was preserved between 0.96 and 0.97.

Following the estimation of the PCM using the information in retained items, final person estimates were also derived. When the two estimates of person measures obtained in STAT 173 data and Lailo's data were compared, a low score, say, 26 had fairly close person estimates (-5.45 logits for STAT 173 and -5.23 logits for Lailo, with a distance of 0.22 logits), whereas a high score, say, 113 had quite far-off transmuted person measures (+3.82 logits for STAT 173 and +2.85 logits for Lailo, with a distance of 0.97 logits). It is evident that the range of student depression measures may have been impacted by sample size or by different temporal characteristics of data. However, it was noted that the relationship between raw score and person measure for both data contexts is approximately linear between raw scores 40 and 110.

To have a glimpse of the prevalence of depression in the given population of college students, separate histograms of person measures for STAT 173 data and Lailo's data were constructed. Figure 6 shows the distribution of students along a continuum of depression severity measures. Although the mean person measure 1.02 (SD=1.74) in STAT 173 data is greater than the mean person measure -0.28 (SD=1.58) in Lailo's data, the shapes of the distributions are very similar, with slight negative skewness, -0.29 (SE=0.13) and -0.56 (SE=0.22), respectively. Despite the different time contexts of populations from which the samples were taken, the consistent prevalence of depression among college students was captured in the analyses of survey data. The STAT 173 data provided a more precise estimate of depression measures since it used a larger number of samples, which reduced the variance observed among person estimates as may be explained by the peaked distribution in Figure 6a.

Figure 6. *Histograms for the student depression measures obtained in the analysis of (a) STAT 173 data and (b) Lailo's data, showing the theoretical normal distribution derived from the mean and standard deviation of person estimates.*



(a) (b)

Distinctive properties of some USDI items were observed in the analysis of Lailo's data. First, three of the 25 remaining items had disordered thresholds, meaning that some response categories did not function as intended. Second, two items did not have 1 responses, meaning that students did not respond "not at all" to these items. These observations were not present in the analysis of STAT 173 data.

Figure 7. *ICCs for 3 USDI items with disordered thresholds in Lailo's Data.*

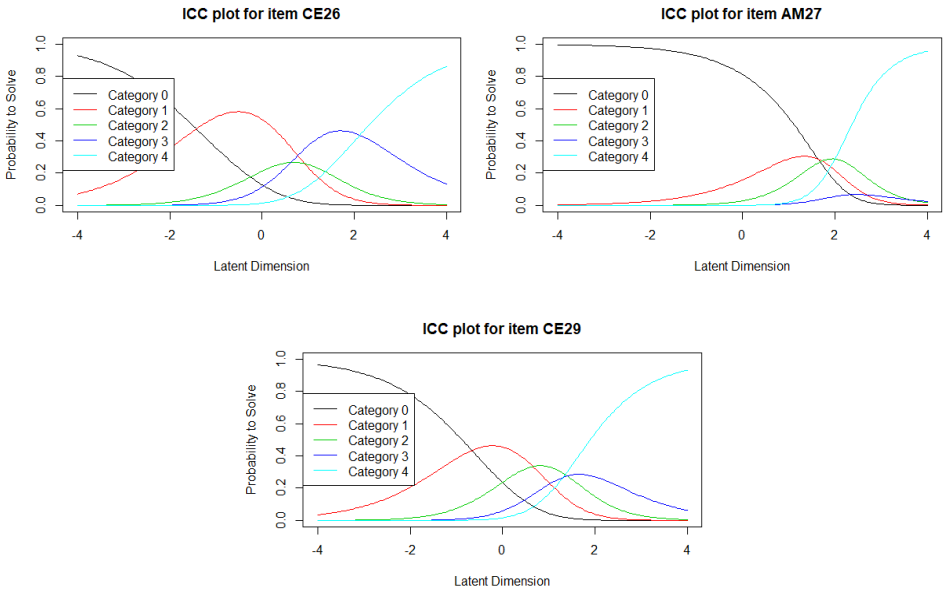


Figure 7 shows the item characteristic curves (ICC; also referred to as item response functions or category characteristics curves) of the three items with disordered thresholds. The threshold estimates for category 2 and 3 in Item CE26 were found to be 0.93 and 0.63, respectively. Also, for items CE29 and AM27, threshold estimates for categories 3 and 4 were also disordered. Items with disordered thresholds disrupt the measurement of the underlying trait (Tennant, 2004).

For these items with response options that did not work as intended, it was suggested that the rating scale structure be revised. For the two items lacking responses in one category, a 4-point response structure (i.e., 1234) be used in future data collection. For the items with disordered

thresholds, the adjacent disordering categories be collapsed into one category (i.e., for CE26 a rating scale structure of 12334; for CE29 and AM27 a rating structure of 12344), hence a 4-point response structure. However, the 5-point scale structure of the rest of the items should be retained.

4. DISCUSSION and CONCLUSION

Based on the results, the procedure applying the extended Rasch model PCM in assessing and streamlining the USDI questionnaire based on two sets of response data yielded comparable results. The scale's items were reduced to the same number for both data contexts, while maintaining instrument reliability. The ordering of the items from various domains (subscales) based on item measures along the continuum of depression severity was found to be consistent with the symptomatology of clinical depression, confirming the construct validity of the streamlined version. Items flagged for LD, DIF and high outfit mean square values were recommended for further investigation of contents, problems in data collection, and possible person subgroup-specific meanings, after which decisions as to revise or entirely discard the items may be made for future use.

The application of the Rasch model in the study was found to be suitable and productive. In addition to the usefulness of outfit and infit statistics in detecting problematic items and persons, other meaningful information about the items and of the entire scale was obtained. The ordering of item severities and distributions of item and person measures provided a basis for the assessment of the instrument's targeting, which is helpful in locating provisional thresholds for various depression severity cases. The detection of item redundancy and bias was also present unlike in traditional item analysis methods. Reliability analysis through person separation provided an evidence of the scale's internal consistency, which is analogous to the Cronbach's alpha. The construction of interval-level measures of student depression would satisfy the conditions set by parametric statistical methods, which makes the computation of effect sizes after implementing interventions and comparison among group means possible.

Overall, the methodology used successfully streamlined the USDI questionnaire, from which person measures were successfully derived. The construction of measures for student depression in both data contexts was also comparable in terms of item threshold estimates. These estimates were used to set provisional thresholds for classifying students of various depression level categories based on combined estimates and, consequently, to help determine the optimal cutoff points when enough data become available. The distributions of student depression measures for both data contexts were found to be consistent despite the difference in time contexts of populations from which these samples were taken. However, estimates for person measures in the two datasets provided varied transmutations for raw scores. The anomalies observed might have been caused by varying sample sizes, survey designs and data collection procedures, items retained in the scale, and temporal characteristics of sample data.

The methodology illustrated in this study explicitly provided a sequence of steps to follow, which is applicable to assessment of other instruments used to measure the prevalence of a latent population characteristic. Generally, there are three steps: (1) fitting the data to the model by eliminating misfits; (2) analyzing retained items; and (3) constructing measures. The sequence was done in the decreasing order of importance. Since the Rasch model requires that data fit the model, decisions on discarding items or persons based on fit statistics have more weights than decisions guided by results of other analyses. In reality, there are solutions for LD and DIF items aside from removal, whereas no solution can be offered to misfitting items/persons but to discard them. It is the strong point of this procedure, espousing the Rasch requirement for invariant measurement, which was empirically demonstrated to be replicable.

In Rasch analysis, the solution for misfits includes the removal of persons with unpredictable responses. However, caution must be taken when ignoring responses from individuals due to model misfit. Another risk to the sample's representativeness is when people responses are eliminated from the survey data. While eliminating outliers improves model fit, a much smaller sample size might result in considerably more serious issues, such as inaccurate estimates of the prevalence of depression and false conclusions. Therefore, in addition to excluding person misfits from analysis of survey data, alternate methods to handle person misfits may be investigated. For example, misfits may be included in the analysis after imputing their health status; this involves replacing the aberrant item response of a person with a given location on the latent continuum by taking into account the responses of good-fit persons with the same location.

It is advised that routine data cleaning be used when analyzing survey data, and that Rasch analysis be used to identify individuals who would unintentionally introduce random noise into subsequent analyses. Rasch analysis cannot pinpoint the specific response bias that may have taken place (e.g., acquiescence, social desirability, guessing, and malingering), but it can at least identify potential sources of measurement noise that could interfere with the identification and estimation of population characteristics, particularly latent traits.

Further studies on the effect of varying proportions of misfits on precision of estimates are encouraged. These studies should include a simulation of Rasch model parameter estimation when person misfits are not included in the data and when they are included to some extent. The estimates obtained in various scenarios can be compared to determine and predict possible impacts of using misfits on the power and validity of survey data. Furthermore, in fitting the Rasch model, these studies should consider the sampling design used in data collection to reduce unwanted bias in the estimation of model parameters. A new statistical package for this purpose can be programmed to facilitate the Rasch analysis of instruments administered in surveys.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Sherwin Balbuena https://orcid.org/0000-0003-0183-4931

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, D.C: American Psychiatric Association.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581-594.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Education Research and Perspectives*, *9*(1), 95-104.

Avery, L.M., Russell, D.J., Raina, P.S., Walter, S.D., & Rosenbaum, P.L. (2003). Rasch analysis of the Gross Motor Function Measure: validating the assumptions of the Rasch model to create an interval-level measure. *Archives of Physical Medicine and Rehabilitation*, *84*(5), 697-705.

Balsamo, M., Giampaglia, G., & Saggino, A. (2014). Building a new Rasch-based self-report inventory of depression. *Neuropsychiatric Disease and Treatment*, *10*, 153.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561-571.

Bond, T.G., & Fox, C.M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Boyle, G.J. (1985). Self-report measures of depression: some psychometric considerations. *British Journal of Clinical Psychology, 24*, 45–59.

Chen, W.H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Deb, S., Banu, P.R., Thomas, S., Vardhan, R.V., Rao, P.T., & Khawaja, N. (2016). Depression among Indian university students and its association with perceived university academic environment, living arrangements and personal issues. *Asian Journal of Psychiatry, 23*, 108-117.

Forkmann, T., Gauggel, S., Spangenberg, L., Brähler, E., & Glaesmer, H. (2013). Dimensional assessment of depressive severity in the elderly general population: Psychometric evaluation of the PHQ-9 using Rasch Analysis. *Journal of Affective Disorders, 148*(2-3), 323-330.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Gesinde, A.M., & Sanu, O.J. (2014). Prevalence and gender difference in self-reported depressive symptomatology among Nigerian university students: Implication for depression counselling. *The Counsellor, 33*(2), 129-140.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stoufer, L. Guttman, E.A. Suchman, P.L. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Habibi, M., Khawaja, N.G., Moradi, S., Dehghani, M., & Fadaei, Z. (2014). University student depression inventory: Measurement model and psychometric properties. *Australian Journal of Psychology, 66*(3), 149-157.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*(1), 56-62.

Hankin, B.L. (2006). Adolescent depression: Description, causes, and interventions. *Epilepsy and Behavior, 8*(1), 102-114.

Hyde, J.S., Mezulis, A.H., & Abramson, L.Y. (2008). The ABCs of depression: Integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression. *Psychological Review, 115*(2), 291-313.

Jeong, H.J., & Lee, W.C. (2016). The level of collapse we are allowed: Comparison of different response scales in Safety Attitudes Questionnaire. *Biometrics and Biostatistics International Journal, 4*(4), 1-7.

Khawaja, N.G., & Bryden, K.J. (2006). The development and psychometric investigation of the University Student Depression Inventory. *Journal of Affective Disorders, 96*(1-2), 21-29.

Khawaja, N.G., Santos, M.L.R., Habibi, M., & Smith, R. (2013). University students' depression: A cross-cultural investigation. *Higher Education Research and Development, 32*(3), 392-406.

Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement, 75*(3): 389-405.

Lailo, J.M.A. (2018). *Determinants of depressive symptoms in undergraduate UPLB students: A joint correspondence analysis*. Institute of Statistics, UPLB.

Lee, R.B., Maria, M.S., Estanislao, S., & Rodriguez, C. (2013). Factors associated with depressive symptoms among Filipino university students. *PloS One, 8*(11): e79825.

Linacre, J.M. (1997). *Guidelines for rating scales MESA Research Note #2*. Available at http://www.rasch.org/rn2.htm.

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, *8*(2), 350.

Lim, G.Y., Tam, W.W., Lu, Y., Ho, C.S., Zhang, M.W., & Ho, R.C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific Reports*, *8*(1), 2861.

Lovibond, P.F., & Lovibond, S.H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335-343.

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20. Available at http://www.jstatsoft.org/v20/i09 .

Maloney, P., Grawitch, M.J., & Barber, L.K. (2011). Strategic item selection to reduce survey length: Reduction in validity? *Consulting Psychology Journal: Practice and Research, 63*, 162-175.

Marcus, M., Yasamy, M.T., Van Ommeren, M., Chisholm, D., & Saxena, S. (2012). *Depression: A Global Public Health Concern*. Geneva: World Health Organization. Available at http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Mikolajczyk, R.T., Maxwell, A.E., El Ansari, W., Naydenova, V., Stock, C., Ilieva, S., ..., & Nagyova, I. (2008). Prevalence of depressive symptoms in university students from Germany, Denmark, Poland and Bulgaria. *Social Psychiatry and Psychiatric Epidemiology*, *43*(2), 105-112.

Nord, M. (2014). *Introduction to Item Response Theory Applied to Food Security Measurement: Basic Concepts, Parameters, and Statistics*. Technical Paper. Rome: FAO. Available at http://www.fao.org/economic/ess/ess-fs/voices/en

O'Connell, M.E., Boat, T., & Warner, K.E. (Eds.). (2009). *Committee on the prevention of mental disorders and substance abuse among children, youth, and young adults: Research advances and promising interventions. Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities.* National Academies Press.

Olsen, L.R., Jensen, D.V., Noerholm, V., Martiny, K., & Bech, P. (2003). The internal and external validity of the Major Depression Inventory in measuring severity of depressive states. *Psychological Medicine*, *33*(2), 351-356.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385-401.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Chapters V-VII, X.

Romaniuk, M., & Khawaja, N.G. (2013). University Student Depression Inventory (USDI): Confirmatory factor analysis and review of psychometric properties. *Journal of Affective Disorders*, *150*(3), 766-775.

Sharif, A.R., Ghazi-Tabatabaei, M., Hejazi, E., Askarabad, M.H., & Dehshiri, G.R. (2011). Confirmatory factor analysis of the University Student Depression Inventory (USDI). *Procedia-Social and Behavioral Sciences*, *30*, 4-9.

Shea, T.L., Tennant, A., & Pallant, J.F. (2009). Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). *BMC Psychiatry*, *9*(1), 1-10.

Smith, R.M. (2000). Fit analysis in latent trait measurement models. *Journal of applied Measurement, 1*(2), 199-218.

Spitzer, R.L., Kroenke, K., Williams, J.B., & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. *JAMA, 282*(18), 1737-1744.

Stansbury, J.P., Ried, L.D., & Velozo, C.A. (2006). Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES–D) scale. *Journal of Personality Assessment, 86*(1), 10-22.

Stanton, J.M., Sinar, E.F., Balzer, W.K., & Smith, P.C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167-193.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tennant, A. (2004). Disordered thresholds: An example from the functional independence measure. *Rasch Measurement Transactions, 17*(4), 945-948

Tennant, A., & Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. *Arthritis Care and Research, 57*(8), 1358-1362.

UPLB INSTAT. (2018). *Utak at Puso: A Survey on the Mental Health Status of UPLB Students (STAT 173 Survey)*. University of the Philippines Los Baños, Laguna.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wongpakaran, N., Wongpakaran, T., & Kuntawong, P. (2019). Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory. *Heliyon*, *5*(8), e02300.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. Geneva: Author. http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1

Wright B.D., & Linacre, J.M. (1987). Dichotomous Rasch model derived from specific objectivity. *Rasch Measurement Transactions, 1*(1), 5-6

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. Chicago, IL: University of Chicago, MESA Press.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*(1), 23-48.

Yu, C.H. (2011). *A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling*. Available at www.creative-wisdom.com/computer/sas/IRT.pdf.

Zigmond, A.S., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370.

Zung, W.W. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*(1), 63-70.

APPENDIX A: UNIVERSITY STUDENT DEPRESSION INVENTORY (USDI) ITEMS WITH FILIPINO TRANSLATION (*in italics*)

In this inventory, the student is asked to indicate how often he/she has experienced each item over the past two weeks by responding in the following 5-point Likert scale: not at all (*hindi kailanman*), rarely (*bihira*), sometimes (*minsan*), most of the time (*madalas*), and all the time (*palagi*).

The prefix of the code indicates the subscale to which the item belongs. Hence, LG, AM, and CE refer to the Lethargy, Academic Motivation, and Cognitive/Emotional subscales, respectively.

Item (*Italics in Filipino*)	Code
1. I am more tired than I used to be. *Ako ay mas pagod ngayon kung ikukumpara sa dati.*	LG01
2. I wonder whether life is worth living. *Napapaisip ako kung may halaga pa bang mabuhay.*	CE02
3. I do not have any desire to go to lectures. *Wala na akong pagnanais na pumasok sa klase.*	AM03
4. I do not have the energy to study at my usual level. *Wala na akong ganang mag-aral gaya ng dati.*	LG04
5. I feel worthless. *Nararamdaman ko na ako ay walang halaga.*	CE05
6. I don't attend lectures as much as I used to. *Mas madalang na ako pumasok sa klase kaysa dati*.	AM06
7. I have thought about killing myself. *Sumagi sa aking isipan na magpakamatay.*	CE07
8. I don't feel motivated to study. *Wala akong motibasyon na mag-aral.*	AM08
9. My energy is low. *Wala akong gana.*	LG09
10. No one cares about me. *Walang nagmamalasakit sa akin.*	CE10
11. I feel emotionally empty. *Wala na akong nararamdamang kahit anong emosyon.*	CE11
12. Going to university is pointless. *Hindi ko nakikita ang kahalagahan ng pagpasok sa unibersidad*	AM12
13. I find it hard to concentrate. *Nahihirapan akong magpokus.*	LG13
14. I feel sad. *Nalulungkot ako*	CE14
15. I worry I will not amount to anything. *Nangangamba akong wala akong mararating sa buhay.*	CE15
16. I don't feel rested even after sleeping.	LG16

Hindi ko ramdam na ako ay nakapagpahinga kahit ako ay nakatulog na.	
17. The activities I used to enjoy no longer interest me. *Nawawalan na ako ng gana sa mga bagay na dating interesado ako.*	CE17
18. Challenges I encounter in my studies overwhelm me. *Nilalamon ako ng mga kinakaharap kong pagsubok sa aking pag-aaral*	LG18
19. I feel like I cannot control my emotions. *Pakiramdam ko, hindi ko na kontrolado ang aking emosyon*	CE19
20. I spend more time alone than I used to. *Napapadalas ang aking pag-iisa.*	CE20
21. My mood affects my ability to carry out assigned tasks. *Nakakaapekto ang aking mga emosyon sa aking abilidad na isagawa ang mga gawaing naiatas sa akin.*	LG21
22. I feel disappointed in myself. *Nakakaramdam ako ng pagkabigo sa aking sarili.*	CE22
23. I have trouble starting assignments. *Nahihirapan akong simulan ang aking mga takdang-aralin.*	AM23
24. Daily tasks take me longer than they used to. *Mas matagal kong maisagawa ang mga pangaraw-araw na gawain kaysa sa nakasanayan.*	LG24
25. I feel withdrawn when I'm around others. *Nakakaramdam ako ng hindi pagkabilang kapag napapaligiran ako ng mga tao.*	CE25
26. I do not cope well. *Hindi na ako makasabay nang maayos.*	CE26
27. I do not find study as interesting as I used to. *Hindi na ako interesadong mag-aral kaysa sa nakasanayan.*	AM27
28. My study is disrupted by distracting thoughts. *Ang aking pag-aaral ay naaantala ng mga nakakaabalang mga saloobin.*	LG28
29. I think most people are better than me. *Sa tingin ko, karamihan sa mga tao ay mas magaling kaysa sa akin.*	CE29
30. I have trouble completing study tasks. *Nahihirapan akong tapusin ang mga gawain ukol sa pag-aaral.*	AM30